AFRL-OSR-VA-TR-2014-0231

SECURE SENSOR SEMANTIC WEB AND INFORMATION FUSION

Bhavani Thuraisingham
UNIVERSITY OF TEXAS AT DALLAS

07/08/2014
Final Report

DISTRIBUTION A: Distribution approved for public release.

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTC
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
06/26/2014

2. REPORT TYPE
Final Report

3. DATES COVERED *(From - To)*
September 2013-May 2014

4. TITLE AND SUBTITLE
Secure Sensor Semantic Web and Information Fusion

5a. CONTRACT NUMBER
FA9550-09-1-0468

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Bhavani Thuraisingham

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of Texas at Dallas,
800 W. Campbell Rd.
Richardson, TX 75080

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Air Force Research Laboratory,
AF Office of Scientific Research,
875 North Randolph St. Rm 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Our objective is to design and develop secure applications to be hosted on the secure cloud infrastructure we have developed under the sister project (Secure Semantic Grid and Cloud). The cloud applications include stream data mining, secure sensor data processing, semantic web data processing, and secure social networks.

15. SUBJECT TERMS
secure cloud, cloud, data mining, secure social networks

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Bhavani Thuraisingham
U	U	U	SAR	29	19b. TELEPHONE NUMBER *(include area code)* 972.883.4738

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Secure Sensor Semantic Web and Information Fusion

Final Report

Contributions by <u>The University of Texas at Dallas</u> in collaboration with

<u>Purdue University</u>
<u>ADB Consulting</u>

Principal Investigator:
Dr. Bhavani Thuraisingham
Bhavani.thuraisingham@utdallas.edu

Period of Performance: September 2013 – May 2014 (as well as some key contributions between 2009 and 2014).

Table of Contents

<u>Contract: FA-9550-09-1-0468</u>

<u>Sponsor: AFOSR</u>

<u>Date: June 25, 2014</u>

1. Introduction

Our objective is to design and develop secure applications to be hosted on the secure cloud infrastructure we have developed under the sister project (Secure Semantic Grid and Cloud). The cloud applications include stream data mining, secure sensor data processing, semantic web data processing, and secure social networks.

During Year 1 of the project, we designed and developed secure graph management for social networks, security preserving data mining for clouds called cloudmask, and information fusion for clouds. These applications are intended to be hosted on the secure cloud infrastructure that includes secure virtual machine monitors, secure storage managers and secure data managers. Figure 1 illustrates our infrastructure.

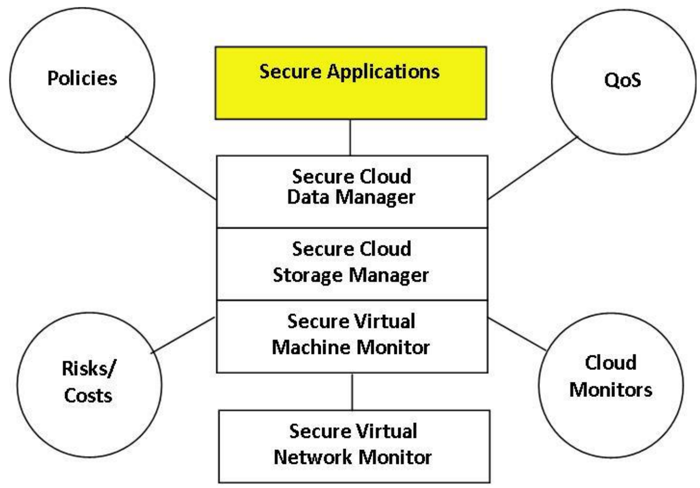

Figure 1. Layered Framework for Assured Cloud

During Year 2 of the project we made substantial contributions in developing a secure social network called VEDANTA. We also worked on adversarial mining in social networks, insider threat detection and trustworthiness of data. In addition, we conducted research on cloud-based assured information sharing and investigated aspects of cyber operations.

During Year 3 we continued with our work on VEDANTA and developed the reasoning component in addition to information extraction. This makes our system unique in that it does information extraction and analysis as well as reasoning for predicting future trends. We also contributed to assured data trustworthiness with Purude, Link Extraction with ADB Consulting, and explored solutions to terrorism problems as well as conducted research on cyber operations. We also conducted research on rumor detection and blocking in social networks that has applications for Department of Defense applications to prevent/detect false information being propagated. Finally we began an investigation with behavioral scientists to study the minds of hackers.

During the final 18 months of the project we investigated ways of transitioning VEDANTA into commercial products and renamed the system InXite. We developed two systems, InXite-Security for security applications and InXite-Marketing, for marketing applications. Multiple patents have been applied for on these systems. We completed a demonstration system for cloud-based assured information sharing. We also wrote a white paper to ONR to transition to technologies to 6-2 research.

Over the course of this project, we have published papers in top tier conferences and journals, published multiple books and gave numerous keynote addresses. The organization of this report is as follows. Our progress will be discussed in Section 2. Our future work will be discussed in Section 3.

2 Progress

Part A: Research

The University of Texas at Dallas

I. A CLOUD-BASED MALWARE DETECTION MODEL FOR EVOLVING MALWARE

Mehedy Masud, Latifur Khan, Kevin Hamlen and Bhavani Thuraisingham
The University of Texas at Dallas

During the past year, we have significantly improved the design of a malware detection module. Here we describe the improvements and their differences from the past works.

Formulating the malware detection problem as data stream classification problem and proposing solution: The problem of detecting malware using data mining involves classifying each executable as either *benign* or *malicious*. Most past methods have approached the problem as a static data classification problem, where the classification model is trained with fixed training data. However, the escalating rate of malware evolution and innovation are not well-suited to static training. Rather, malware detection now should be treated as a *data stream* classification problem. In this scenario, the data stream is a sequence of executables in which each data point is one executable. The stream is *infinite-length*. It also observes *concept-drift* as attackers relentlessly develop new techniques to avoid detection, changing the characteristics of the malicious code. Similarly, the characteristics of benign executables change with the evolution of compilers and operating systems. Therefore, in our current work, we have approached the malware detection problem as a data stream classification problem.

We propose a multi-partition, multi-chunk ensemble classification algorithm that generalizes existing ensemble based data stream classification techniques. The generalization leads to significantly improved classification accuracy relative to existing single-partition, single-chunk ensemble approaches when tested on real-world data streams. Our approach divides the data stream into equal sized *chunks*. The chunk size is chosen so that all data in each chunk fits into the main memory. Each chunk, when *labeled*, is used to train classifiers. The approach is therefore parameterized by the number of partitions v, the number of chunks r, and the ensemble size K. An ensemble consists of Kv classifiers. Whenever a new data chunk is labeled, the ensemble is updated. We take the most recent labeled r consecutive data chunks and train v classifiers using v-fold partitioning of these chunks. We then update the ensemble by choosing the best Kv classifiers (based on accuracy) among the newly trained v classifiers and the existing Kv classifiers. Thus, the total number of classifiers in the ensemble remains constant, which addresses the infinite length problem of data streams. Furthermore, by keeping the ensemble updated, the concept-drift problem is addressed.

Using cloud for feature extraction and feature selection Feature extraction and selection is an important part in building a data mining model. In the current data stream setting, this part is even more important because it is the main bottleneck in building the data mining model. From our past experience with binary n-gram feature extraction from executables, we observe that for a training corpus of only 3500 executables, extraction and selection required about 2 hours of computation and many gigabytes of disk space for a machine with a quad-core processor and 12GB of memory. In the current work, we consider a much larger dataset of 105 thousand executables for which our previous approach is insufficient. We therefore propose a scalable feature selection and extraction solution that leverages a cloud computing framework. We show that depending on the availability of cluster nodes, the running time for feature extraction and selection can be reduced by a factor of m, where m is the number of nodes in the cloud cluster. The nodes are machines with cheap commodity hardware. Therefore, the solution is also cost-effective as high-end computing machines are not required.

Addressing the dynamic feature space problem Due to the dynamic nature of the stream, each chunk of data may have a different best set of features. In our case, we also observe this scenario and found that each data chunk in the stream has different set of best (say *K*) features. This creates a problem in classification because of the heterogeneity in feature space between the classification model and the test instances. We have proposed a solution to this problem by taking the "union" of the feature spaces, which shows significant improvement in classification accuracy over previous approaches that use either fixed feature set, or "intersection" of feature sets.

II. ADVERSARIAL DATA MINING: THEORY AND APPLICATIONS

Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham
The University of Texas at Dallas

We tackle data mining problems in an adversarial environment where an adversary attempts to influence statistical analyses by modifying data. Our efforts manifest their intended results in three major publications on both theory and applications in adversarial learning. The three research projects include:

- Randomizing smartphone malware profiles against statistical mining techniques;
- Adversarial support vector machine learning;
- Sparse Bayesian Adversarial Learning Using Relevance Vector Machine Ensembles.

Randomizing Smartphone Malware Profiles against Statistical Mining Techniques
Compared to conventional mobile phones, smartphones are built to support more advanced computing needs and meet modern custom software demands. An unpleasant byproduct of the ongoing smartphone revolution is its invitation to malicious exploits. As smartphone software grows more complex, more malware programs will be created to attempt to exploit specific weaknesses in smartphone software. Smartphones of end users all together constitute a large portion of the powerful mobile network. Having access to the enormous amount of personal information on this network is a great incentive for the adversary to attack the smartphone mobile world.

Malicious activities on mobile phones are often carried out through lightweight applications, scrupulously avoiding detection while leaving little trace for malware analysis. Over the years many malware detection techniques have been proposed. These techniques can be roughly divided into two groups: static analysis and dynamic analysis. Static analysis techniques discover implications of unusual program activities directly from the source code. Although static analysis is a critical component in program analysis, its ability to cope with highly dynamic malware is unsatisfactory. A number of obfuscation techniques have been shown to easily foil techniques that rely solely on static analysis. Dynamic analysis (also known as behavioral analysis) identifies security holes by executing a program and closely monitoring its activities. Information such as system calls, network access, and files and memory modifications is collected from the operating system at runtime. Since the actual behavior of a program is monitored, threats from dynamic tactics such as obfuscation are not as severe in dynamic analysis. However, dynamic analysis cannot guarantee a malicious payload is always activated every time the host program is executed.

We follow a similar perspective of dynamic analysis by analyzing real-time collections of statistics of smartphone usage. Metrics of real-time usage are recorded for analysis. We choose the Android platform in our study. Android is open source and apparently has a solid customer base given that many devices are using this platform. For the convenience of security analysis on this platform, we developed custom parameterized malware programs on the Android platform. These malware programs can target the victim for the purpose of denial of service attacks, information stealing, and privacy intrusion. Our second contribution is the empirical analysis of the weaknesses of data mining techniques against mobile malware. We demonstrate that a malware program with unpredictable attacking strategies is more resilient to commonly used data mining techniques.

Adversarial support vector machine learning

Many learning tasks, such as intrusion detection and spam filtering, face adversarial attacks. Adversarial exploits create additional challenges to existing learning paradigms. Generalization of a learning model over future data cannot be achieved under the assumption that current and future data share identical properties, which is essential to the traditional approaches. In the presence of active adversaries, data used for training in a learning system is unlikely to represent future data the system would observe. The difference is not just simple random noise which most learning algorithms have already taken into consideration when they are designed. What typically flunk these learning algorithms are targeted attacks that aim to make the learning system dysfunctional by disguising malicious data that otherwise would be detected. Existing learning algorithms cannot be easily tailored to counter this kind of attack because there is a great deal of uncertainty in terms of how much the attacks would affect the structure of the sample space. Despite the sample size and distribution of malicious data given at training time, we would need to make an educated guess about how much the malicious data would change, as sophisticated attackers adapt quickly to evade detection. Attack models, that foretell how far an adversary would go in order to breach the system, need to be incorporated into learning algorithms to build a robust decision surface. In this paper, we present two attack models that cover a wide range of attacks tailored to match the adversary's motives. Each attack model makes a simple and realistic assumption on what is known to the adversary. Optimal SVM learning strategies are then derived against the attack models.

Some earlier work lays important theoretical foundations for problems in adversarial learning. However, earlier work often makes strong assumptions such as unlimited computing resource and both sides having a complete knowledge of their opponents. Some propose attack models that may not permit changes made to arbitrary sets of features. In security applications, some existing research mainly explores practical means of defeating learning algorithms used in a given application domain.

Meanwhile, various learning strategies are proposed to fix application-specific weaknesses in learning algorithms, but only to find new doors open for future attacks. The main challenge remains as attackers continually exploit unknown weaknesses of a learning system. Regardless of how well-designed a learning system appears to be, there are always "blind" spots it fails to detect, leading to escalating threats as the technical strengths on both sides develop. Threats are often divided into two groups, with one group aiming to smuggle malicious content past a learning-based detection mechanism, while the other tries to undermine the credibility of a learning system by raising both false positive and false negative rates. The grey area in between is scarcely researched. In this work, we set ourselves free from handling application-specific attacks and addressing specific weaknesses of a learning algorithm. Our main contributions lie in the following three aspects:

- We develop a learning strategy that solves a general convex optimization problem where the strength of the constraints is tied to the strength of attacks.
- We derive optimal support vector machine learning models against an adversary whose attack strategy is defined under a general and reasonable assumption.
- We investigate how the performance of the resulting optimal solutions changes with different parameter values in two different attack models. The empirical results suggest our proposed adversarial SVM learning algorithms are quite robust against various degrees of attacks.

Sparse Bayesian Adversarial Learning Using Relevance Vector Machine Ensembles

Existing research in adversarial learning varies in the types of constraints considered in the problem definition. The assumption of unconstrained adversaries is impractical since arbitrary modification to data and its class membership can result in a worst-case error rate of 100%. Therefore, the majority of the recent research focuses on constrained adversaries. Under the constrained-adversary assumption, major research results can be further divided between game-theoretic solutions and non-game theoretic solutions. For practitioners, the difficulty lies in choosing the most appropriate method for problems at hand. Solutions developed in the game-theoretic framework almost always assume a rational game. In

addition, each player is assumed to have a certain amount of knowledge about the opponent. Similarly, non-game theoretic methods often make assumptions on the opponent's knowledge, the distribution of corrupted data, and available computing resources. In practice, adversaries are seldom optimal and the knowledge and the resources they possess are hard to assess.
For classification problems, the common assumption is that data are independently and identically distributed. This assumption is easily violated when there is an active adversary who modifies data to influence the prediction. When data is constantly modified in an unpredictable way, training data would never be sufficient to induce an accurate classifier. On the positive side, at training time we can explore the feature space and find the most effective direction for the adversary to move data in the feature space to influence the classifier. Once we find such a direction, we can improve the classifier by countering these potential moves. The learning model we choose to implement this strategy is the relevance vector machine.

Similar to the support vector machine method, the relevance vector machine (RVM) is a sparse linearly parameterized model. It is built on a Bayesian framework of the sparse model. Unlike the support vector machine in which a penalty term is introduced to avoid over-fitting the model parameters, the relevance vector machine model introduces a prior over the weights in the form of a set of hyperparameters, one associated independently with each weight. Very large values of the hyperparameters (corresponding to zero- weights) imply irrelevant inputs. Training data points associated with the remaining non-zero weights are referred to as relevance vectors. The relevance vector machine typically uses much fewer kernel functions compared to the SVM.

In this paper, we propose a sparse relevance vector machine ensemble for adversarial learning. The basic idea of this approach is to learn an individual kernel parameter η_i for each dimension di in the input space. The parameters are iteratively estimated from the data along with the weights and the hyperparameters associated with the weights. The kernel parameters are updated in each iteration so that the likelihood of the positive (malicious) data points are minimized. This essentially models an adversarial attack as if the adversary were granted access to the internal states of the learning algorithm. Instead of using fixed kernel parameters, we search for kernel parameters that simulate worst-case attacks while the learning algorithm is updating the weights and the weight priors of a relevance vector machine. We learn M such models and combine them to form the final hypothesis. Our main contributions are:

- Extending the sparse Bayesian relevance vector machine model to counter adversarial attacks;
- Developing a kernel parameter fitting technique to model adversarial attacks within the RVM framework.

The use of individualized kernel parameters has been shown beneficial to kernel-based learning; however, this is the first time it is applied to adversarial learning.

III. DATA ANALYTICS FOR INSIDER THREAT DETECTION

Parveen Pallabi and Bhavani Thuraisingham
The University of Texas at Dallas

Evidence of malicious insider activity is often buried within large data streams, such as system logs accumulated over months or years. Ensemble-based stream mining leverages multiple classification models to achieve highly accurate anomaly detection in such streams, even when the stream is unbounded, evolving, and unlabeled. This makes the approach effective for identifying insiders who attempt to conceal their activities by varying their behaviors over time.

In our approach we have applied ensemble-based stream mining, supervised and unsupervised learning, and graph-based anomaly detection to the problem of insider threat detection. It demonstrates that the ensemble-based approach is significantly more effective than traditional single-model methods,

supervised learning outperforms unsupervised learning, and increasing the cost of false negatives correlates to higher accuracy. It shows effectiveness over non-sequence data.

For sequence data, we designed, developed and tested an unsupervised, ensemble-based learning algorithm that maintains a compressed dictionary of repetitive sequences found throughout dynamic data streams of unbounded length to identify anomalies. In unsupervised learning, compression-based techniques are used to model common behavior sequences. This results in a classifier exhibiting a substantial increase in classification accuracy for data streams containing insider threat anomalies. This ensemble of classifiers allows the unsupervised approach to outperform traditional static learning approaches and boosts the effectiveness over supervised learning approaches. One of the bottlenecks to construct compress dictionary is scalability. For this, an efficient solution is proposed and implemented using a Hadoop and MapReduce framework.

We have several publications, a PhD thesis and a book contract signed on "Big Data Analytics with Applications in Insider Threat Detection".

IV. SECURE DATA PROVENANCE AND ACCESS CONTROL

Tyrone Cadenhead, Vaibhav Khadilkar, Murat Kantarcioglu, Bhavani Thuraisingham
The University of Texas at Dallas

Inference is the process of forming conclusions from premises. The inferred knowledge is harmful when the user is not authorized to acquire such information from legitimate responses that he/she receives. Providing a solution to the inference problem where users issue multiple requests, and consequently infer unauthorized knowledge is an open problem. An inference controller is a device that is used to detect or prevent the occurrence of the inference problem. However, an inference controller will never know, in full, the inferences possible from the answers to a query request, since there is always some prior knowledge available to the querying user. This prior knowledge could be any subset of all possible knowledge available from other external sources. The inference problem is complex and therefore, an integrated and/or incremental domain specific approach is necessary for its management. For a particular domain, one could take several approaches, such as: 1) building inference controllers which act during query processing; 2) building inference controllers which enforce constraints during the knowledge base design; and 3) building inference controllers, which provide explanations to a system security officer. This book discusses the implementation of these incremental approaches for a prototype inference controller for provenance in a medical domain.

Provenance is metadata that captures the origin of a data source; the history or ownership of a valued object or a work of art or literature. It allows us to verify the quality of information in a data store, to repeat manipulation steps and to discover dependencies among data items in a data store. In addition, provenance can be used to determine the usefulness and trustworthiness of shared information. The utility of shared information relies on: (i) the quality of the source of information; and (ii) the reliability and accuracy of the mechanisms (i.e., procedures and algorithms) used at each step of the modification (or transformation) of the underlying data items. Furthermore, provenance is a key component for the verification and correctness of a data item, which is usually stored and then shared with information users. We have designed and developed an inference controller that operates over provenance. This controller protects the sensitive information in a provenance database from unauthorized users. The provenance is represented as a directed acyclic graph. This graphical representation of provenance can be represented and stored using semantic web technologies. We have built a prototype to evaluate the effectiveness of our inference controller. We store the provenance in a semantic web-based knowledge base and use semantic web reasoners to draw inferences from the explicit information in a provenance graph. We enforce constraints at the design phase as well as at runtime.

Our work on secure data provenance has developed flexible and scalable access control policies by extending role-based access control (RBAC) using key semantic web technologies. We also implemented a prototype, which shows that we can scale and reason over a set of access control policies efficiently. We provided a definition of an access control policy language for provenance. This language retains the properties of traditional access control to gain access to data. Furthermore, the language provides an additional advantage whereby we can write one policy which is a pattern for several policies, thus contracting the policy set. We also build a prototype using semantic web technologies that allows a user to query for data and provenance based on access control policies defined using our policy language. We investigated the application of a graph grammar technique, which can be used to perform redaction over provenance. In addition, we developed an architectural design that allows a high-level specification of policies, thus separating the business layer from a specific software implementation. We also implemented a prototype of the architecture based on open source semantic web technologies. With respect to inference control, we designed an inference architecture, which uses a risk-based model to determine whether provenance can be released. In particular, we developed a query processing approach for inference control with provenance data based on query modification with SPARQL.

V. CLOUD-BASED ASSURED INFORMATION SHARING

Tyrone Cadenhead, Vaibhav Khadilkar, Murat Kantarcioglu, Bhavani Thuraisingham
The University of Texas at Dallas

The advent of *cloud computing* and the continuing movement toward *software as a service* (SaaS) paradigms have posed an increasing need for *assured information sharing* (AIS) as a service in the cloud. The urgency of this need has been voiced as recently as April 2011 by NSA (National Security Agency) CIO (Chief Information Officer) Lonny Anderson in describing the agency's focus on a "cloud-centric" approach to information sharing with other agencies. Likewise, the DoD (Department of Defense) has been embracing cloud computing paradigms to more efficiently, economically, flexibly, and scalably meet its vision of "delivering the power of information to ensure mission success through an agile enterprise with freedom of maneuverability across the information environment". Both agencies therefore have a tremendous need for effective AIS technologies and tools for cloud environments.

Although a number of AIS tools have been developed over the past five years for policy-based information sharing, to our knowledge none of these tools operate in the cloud and hence do not provide the scalability needed to support large numbers of users utilizing massive amounts of data. Our recent prototype systems for supporting cloud-based AIS have applied cloud-centric engines that query large amounts of data in relational databases via non-cloud policy engines that enforce policies expressed in XACML. While this is a significant improvement over prior efforts (and has given us insights into implementing cloud-based solutions), it nevertheless has at least three significant limitations. First, XACML-based policy specifications are not expressive enough to support many of the complex policies needed for AIS missions like those of the NSA and DoD. Second, to meet the scalability and efficiency requirements of mission-critical tasks, the policy engine needs to operate in the cloud. Third, secure query processing based on relational technology has limitations in representing and processing unstructured data needed for many applications.

To share the large amounts of data securely and efficiently, there clearly needs to be a seamless integration of the policy and data managers in the cloud. Therefore, in order to satisfy the cloud-centric AIS needs, we need (i) a cloud-resident policy manager that enforces information sharing policies expressed in a semantically rich language, and (ii) a cloud-resident data manager that securely stores and retrieves data and seamlessly integrates with the policy manager. To our knowledge, no such system currently exists. Therefore, our project to design and develop such cloud-based assured information sharing system proceeded in two phases.

We have designed a system and implemented a version a Cloud-centric Assured Information Sharing System (CAISS) that utilizes the technology components we have designed in-house as well as some open source tools. CAISS consists of two components: a cloud-centric policy manager that enforces policies specified in RDF (resource description framework), and a cloud-centric data manager that will store and manage data also specified in RDF. This RDF data manager is essentially a query engine for SPARQL (SPARQL Protocol and RDF Query Language), a language widely used by the semantic web community to query RDF data. RDF is a semantic web language that is considerably more expressive than XACML for specifying and reasoning about policies. Furthermore, our policy manager and data manager will have seamless integration since they both manage RDF data. We have chosen this RDF-based approach for cloud-centric AIS during Phase 1 because we have already developed an RDF-based non-cloud centric policy manager and an RDF-based cloud-centric data manager. Specifically, we enhanced our RDF-based policy engine to operate on a cloud, extended our cloud-centric RDF data manager to integrate with the policy manager, and built an integrated framework for CAISS.

While our CAISS design and implementation was the first system supporting cloud-centric AIS, it operates only on a single-trusted cloud and will therefore not support information sharing across multiple clouds. Furthermore, while CAISS's RDF-based, formal semantics approach to policy specification will be significantly more expressive than XACML-based approaches, it will not support an enhanced machine interpretability of content since RDF does not provide a sufficiently rich vocabulary (e.g., support for classes and properties). We have therefore designed a fully functional and robust AIS system called CAISS++ that addresses these deficiencies.

VI. REDACT: A FRAMEWORK FOR SANITIZING RDF

Jyothsna Rachapalli, Vaibhav Khadilkar, Murat Kantarcioglu, Bhavani Thuraisingham
The University of Texas at Dallas

RDF data sanitization is the process of masking sensitive data in an RDF graph with a suitable replacement in order to mitigate the risk of data exposure. RDF sanitization can be useful under two usage scenarios. Firstly, when an RDF dataset needs to be outsourced or shared with a third party, in which case sanitization can be performed on the entire dataset before being shared. Secondly, in an access control-like scenario where data is present in original/unmasked form but hidden from those who do not have access to it. In this case, sanitization is performed on a per query basis on the subgraph of the RDF dataset that is being accessed by the user query. *Why sanitize RDF data?* An organization must take measures to secure its data from accidental exposure or malicious exposure by an insider, as data privacy and security have become drivers for maintaining brand, reputation and customer satisfaction. Moreover, organizations must also comply with federal and state regulations surrounding information security and privacy, through laws such as HIPAA, Data Protection Act, etc.

Since security or privacy definitions and intent change by virtue of application requirements that span a vast number of domains, we need a general and fundamental mechanism for securing RDF graphs, which should essentially comprise RDF graph transformation operations. Towards this end, we designed and developed **REDACT** (**R**df **ED**iting **A**nd **C**oncealing **T**ool), which comprises a set of fundamental RDF graph sanitization operations that are built as an extension to SPARQL. These operations can be used to transform, manipulate or sanitize RDF graphs by concealing sensitive data. *A motivating scenario: Assured Information Sharing (AIS)* refers to organizations sharing information while enforcing policies and procedures so that the integrated data can be queried securely to extract nuggets. An AIS system integrating data sources from the Army, Navy, Air Force, Local, State and Federal agencies as well as medical databases is critical, as it may contain sensitive information about secret service agents, potential terrorists, etc. The data sources are integrated so that the big picture thus formed can be queried, patterns

and information extracted and time sensitive crucial decisions made by informing all concerned parties. While the different agencies have to share data, they need to do so in a secure manner by sanitizing the sensitive data. RDF provides an elegant solution for the data integration needs of the above scenario as it was fundamentally designed for such purposes. However, it falls short in providing a means for sanitization. *What Data to Sanitize (Problem Scope)?* We choose to leave this decision to application developers as it is a domain specific issue. We instead try to address *How to sanitize RDF data?* by providing a set of RDF sanitization operations, which can be suitably used by application developers to sanitize and protect the RDF data that they consider to be sensitive.

Our contributions:

- We have extended SPARQL through a set of fundamental graph sanitization operations to secure RDF data. In addition, we also present the time complexity analysis of these operations.
- We have developed denotational semantics of this extension of SPARQL.
- We have developed a prototype system and its architecture based on a healthcare provenance scenario and illustrate how one can build more complex security features using our graph sanitization operations.
- We have obtained empirical results showing the performance of the sanitization operations, which were evaluated on synthetic as well as real world datasets.

VII. RUMOR BLOCKING IN SOCIAL NETWORKS

Lidan Fan, D. Z. Du, Bhavani Thuraisingham
The University of Texas at Dallas

It is a common phenomenon that some individuals may spread a rumor about a person, or a product of a company, or a service of a company, etc. Due to the special structure of social networks, especially online social networks, such as Facebook and Twitter, information is able to spread very fast and reaches a large number of people within seconds. Thus, when a rumor or misinformation spreads in social networks, it may cause terrible results among the public.

The objective of our work is to design efficient algorithms to block rumor propagation. One strategy we adopt is to launch the opposite cascade of rumors, called protectors, to fight against rumors. We study the Rumor Control problem: protect as many individuals as possible in a social network by selecting the least number of individuals as initial protectors. Considering the community property of social networks, we define a special kind of vertex set, called bridge ends. They play the roles as gates for their own communities. The goal of the problem is to protect certain fraction of the bridge ends with minimal number of initial protectors. We propose an influence propagation model: the Deterministic One-Activate-Many model. We prove that there is no polynomial time $o(\ln n)$-approximation for it unless $P = NP$ and design a Set Cover Based Greedy algorithm with $O(\ln n)$-approximation ratio, where n is the number of bridge ends. Finally, we compare our algorithm with several heuristic algorithms in two datasets obtained from real world, and the results demonstrate that our algorithm outperforms those heuristic algorithms.

We also investigated the problem: given the number of initial protectors and deadline, how to select initial protectors such that the number of "really" protected members in social networks is maximized within deadline. We propose two models, namely the Rumor-Protector Independent Cascade model with Meeting events model and the Rumor-Protector Linear Threshold model with Meeting events model, to capture both rumor and protector diffusion processes. Three features are included in these two models: a time deadline, random time delay between information exchange and personal interests regarding the acceptance of information. Under these two models, we study the RC-DMP problem. We prove that the problem under these two models is NP-hard. Moreover, we demonstrate that the objective functions for

this problem under the two different models are both monotone and submodular. Therefore, we apply a greedy algorithm as a constant-factor approximation algorithm with a performance ratio of $1-\frac{1}{e}$.

We also addressed the rumor blocking in cellular networks, where we consider a mobile worm as a rumor. Our strategy is to distribute patches to some "effective" nodes in a cellular network. Our goal is to choose minimal number of "effective" nodes, such that all the nodes in this network can be protected within one time step before rumor (mobile worm) propagation. We propose a novel influence diffusion model: the Asymmetric-Trust Infection model, which incorporates each individual's trust towards their friends. We analyze the complexity of the problem and prove the objective function for the problem is a polymatroid function. Therefore, we present a Greedy Algorithm, which has an approximation solution with a factor of $1 + 2\ln\delta$ from optimal, where δ is the maximum degree of the input graph.

VIII. SECURE SOCIAL NETWORK: VEDANTA

Satyen Abrol, Vaibhav Khadilkar, Latifur Khan, Bhavani Thuraisingham
The University of Texas at Dallas

Since its inception in the mid-90s, social networks have provided for a way for users to interact, reflecting of social networks or social relations among people, e.g., who share interests and/or activities. At the forefront of emerging trends in social networking sites is the concept of "real time" and "location based". So what makes location based social media services so important?

Privacy and Safety: Posting updates on location-based social networking websites and publishing your current location to the user can result in problems like personalized attacks by spammers, and threats to your safety.

Trustworthiness of User Location: In certain scenarios, such as the political scenario of the Iran elections of 2009, it becomes important for organizations monitoring the data to be able to verify the location of a user.
Advertising and Marketing: Social networks connect people at low cost and can be beneficial for entrepreneurs and small businesses looking to expand their contact bases. These networks often act as a customer relationship management tool for companies selling products and services.

Having highlighted the importance of the location of the user in social media, it is important to understand that it is not provided explicitly by the users for a number of reasons. Some of the users are concerned about their privacy and security; others do not find any incentive in sharing the location. Apart from this class of users who do not disclose their location, there are others who provide locations which are either incorrect or not machine readable or reveal just the state/country. The unstructured and free form of the text consisting of internet slang and incomplete sentences makes use of traditional Natural Language Processing and gazetteer-based data mining approaches produce inaccurate results.

We develop a social intelligence application, Vedanta, to identify the location of the user on social networking sites by mining information from his social graph and the messages posted by him. Vedanta not only identifies the city level home location of a user, but goes one step further to pinpoint specific venues or point of interests that the user may have visited or has talked about in his messages. We have performed extensive experiments to prove the efficacy of the algorithm in terms of accuracy and running time. The algorithm outperforms all existing location extraction approaches. To show the applicability of the algorithm in security analytics, we developed a powerful tool that allows analysts to identify the location of any *Twitter* user and his friends, tie in text to reveal what different users are talking about around the world in real time. The tool provides an intuitive graphical interface which an analyst can use to visualize the places visited by a user and his friends (determined by the algorithm) to identify and monitor potential security threats. In the present world scenario where uprisings, political meetings such

as the Arab spring or London riots are organized on social networking sites like *Twitter*, Vedanta proves to be a great tool for detecting, recognizing and tracking users with mal-intent.

Much of the focus during previous years was on developing information integration as well as the reasoning components. In particular, information from multiple social networks was integrated securely and analyzed and reasoned to detect future events. Our work was presented also at various conferences. The focus during the last 18 months has been to develop more robust applications for security and marketing.

IX. CYBER OPERATIONS

Jan Kallberg and Bhavani Thuraisingham
The University of Texas at Dallas

The Principal Investigator has a Certificate in Terrorism Studies from St. Andrews University in Scotland. Therefore as part of this project she is working with research scientists who have strong expertise in this area to identify major problems faced by our nation today and develop solutions. During Year 2 we wrote a series of papers on terrorism studies partially funded by this project. Our first paper is on tracking Al-Qaeda financial networks. The goal is to thwart the financial networks to weaken the terrorists. Our second paper is in studying the financial crisis of 2007-2009. It is critical that we are secure financially so that sufficient funds can be allocated to fund terrorism related activities. The financial crisis of 2007-2009 has put us in a weaker position. Therefore, we need to analyze the root of the problem of the financial crisis to ensure that this does not happen in the future. Our third paper is on credit card fraud analysis. While there has been work in this area, we are giving our views about this topic. Our fourth paper, is on financial systems security. The first two papers have been presented at the IEEE Intelligence and Security Informatics Conference (ISI) and European ISI Conference. A book chapter has been prepared. During Year 3 we investigated Cyber Operations. We have published a series of papers on this topic. A book contract has also been signed on cyber operations. In addition we have begun collaboration with behavioral scientists to study the mind of hackers and develop systems based on this study. For the past eighteen months we have continued with our work on cyber operations and presented the work to IBM. We received an IBM Faculty Award to prepare courses in this area.

We have successfully been published in several Department of Defense (DoD) journals: *Joint Forces Quarterly, Strategic Studies Quarterly, Air- and Space Journal, and Military Review.* We have addressed several innovative approaches to cyber defense and cyber operations. Through our research we have found several unique tenets of cyber that are providing a foundation for future research. Earlier works have transposed traditional military strategy into a cyber context, but our research serves several angles where traditional military theory either struggles or fails. Briefly, they are categorized as anonymity, object permanence, measurement of success, and time window. Military theorists are mainly addressing battling a known entity who exists in space and time, in a conflict where you can assess battle damage and effectiveness, and the conflict occurs within a time frame that allows leadership to act. The result of our research questions the fitness of these strategies in cyber. Especially concerning measurement of effectiveness, cyber lacks a feedback loop to tell if a counter strike was effective or not, and the short time frame in which future digital interchanges will occur.

We were invited to USMA West Point to give feedback and provide guidance in the creation of an Army Cyber Institute based on our ability to address cyber operations in a wider and societal perspective. An additional manuscript was submitted at the end of the reporting period to Joint Forces Quarterly titled: *Strategic Cyberwar Theory - A Foundation for Designing Decisive Strategic Cyber Operations*. According to Colonel Greg Conti, head of Army Cyber Institute, it presents a road map to bring down an adversarial nation, which we consider validates that we are on a viable track in our cyber operations research.

X. BEHAVIORAL ANALYSIS

Daniel Krawczyk and James Bartlett, Murat Kantarcioglu, Bhavani Thuraisingham
The University of Texas at Dallas

Toward the ultimate goal of enhancing human performance in cyber security, we attempt to understand the cognitive components of cyber security expertise. Our initial focus is on cyber security attackers – often called “hackers”. Our first aim is to develop behavioral measures of accuracy and response time to examine the cognitive processes of pattern-recognition, reasoning and decision-making that underlie the detection and exploitation of security vulnerabilities. Understanding these processes at a cognitive level will lead to theory development addressing questions about how cyber security expertise can be identified, quantified, and trained. In addition to behavioral measures, our plan is to conduct a functional magnetic resonance imaging (fMRI) study of neural processing patterns that can differentiate persons with different levels of cyber security expertise. Our second aim is to quantitatively assess the impact of attackers’ thinking strategies – conceptualized by psychologists as heuristics and biases – on their susceptibility to defensive techniques (e.g., “decoys,” “honeypots”). Honeypots are an established method to lure attackers into exploiting a dummy system containing misleading or false content, distracting their attention from genuinely sensitive information, and consuming their limited time and resources. We use the extensive research and experimentation that we have carried out to study the minds of successful chess players in order to study the minds of hackers with the ultimate goal of enhancing the security of current systems. We prepared an exploratory paper expressing our ideas.

Subcontractors

XI. IMPROVING DATA TRUSTWORTHINESS

Elisa Bertino, Purdue University

Murat Kantarcioglu, Collaborator , The University of Texas at Dallas

XI.A. ENHANCING DATA TRUSTWORTHINESS IN SOCIAL NETWORKS

The work has addressed the problem of data trustworthiness for social networks. Social networks have been studied by various research communities for more than fifty years. However, the advent of the online social networks and the wide adoption of such networks have significantly increased the importance of obtaining useful information from those networks. Extracting useful knowledge from social network datasets proves to be a difficult problem and social network mining is currently identified as one of the most challenging problems in data mining research. To add to the difficulty of this problem, privacy concerns exist for many social network datasets. Such concerns have resulted in limited accessibility to social network data and thus in reducing the quantity and quality of the knowledge that could be extracted from these datasets. Such knowledge may have important applications, such as disease spreading in epidemiology, emergency management, protection from cyber attacks, etc.

While large online social networks such as Facebook and LinkedIn are well-known and gather millions of users, small social networks are today becoming increasingly common. Currently, such small niche social networks such as GoFISHn and GoHUNTn are considered as the new trend in online social network usage. Many organizations already use existing social networks to connect to their customers and users. Seeing the increasing usage of small social networks, such organizations will likely start to create in-house online social networks where they will own the data shared by customers. Nowadays, for many services (insurance, airline miles, travel sites, etc.), users have individual accounts on organizations websites. However, there is no network structure connecting accounts of different users, and therefore the relationships that may exist among such users are not efficiently used by organizations. The benefits that can be obtained from adding relationships among users are significant in order to enhance the knowledge that can be extracted by social network data. A challenge is that users must have an incentive to connect among themselves in an organization-owned social network. This is not a trivial problem and will likely be a difficult challenge to address. However, the use of incentives will motivate the users in connecting to their friends or acquaintances. For instance, an insurance company may use incentives such as 10% savings on their car insurance costs if a customer registers on its social network site and recommends a minimum number of friends. Next, the amount of savings can be increased based on how many of his/her friends will buy insurance from the same company. Such incentives could also be used to motivate a user to complete his/her profile, and this would allow the insurance company to have a wealth of information about its users that could potentially be used to increase its business.

It can be easily seen that such local social networks have many benefits for the organizations that own them. However, the users' main motivation for joining and providing the required information is to get the desired service at a discount price or any other incentive associated with the use of this organization-owned social networking site. Therefore, it is expected that users will be less likely to provide only accurate information in their profiles (due to privacy concerns or because of other advantages that could be obtained by partially faking profile information or for other reasons, like constructing fake online identities and creating online connections with specific individuals in order to target them for spear-phishing attacks). An example of a possible advantage that could be obtained is as follows. A user can report his marital status as single although he is married. The reason of such reporting is that his wife may be under 25 years old and adding her in the profile may result in the insurance agency including her in the insurance policy and therefore increasing the auto insurance rate. Other examples include misreporting of address, age, and so on. This possibility of faking part of profile data will diminish the utility of the data. The organizations that own such data will benefit from it if they can assess the trustworthiness of such data and can identify possible fake information. Unfortunately, due to privacy regulations, large social

network datasets that could potentially be used to verify local information may not be available due to privacy concerns. However, we can expect that anonymized social network datasets be available and they can be used to determine the trustworthiness of local data.

The work has investigated the problem of data trustworthiness in social networks when repositories of anonymized social networks exist and has designed and validated approaches by which one can assign a *trust score* to user profiles (or specific information within user profiles) in a social network. The trust score is a numeric indicator ranging from 0 to 1 that conveys the confidence that the associated information is truthful; a value close to 0 indicates a low confidence whereas a value closer to 1 indicates high confidence. Notice that this trust score is just an indicator and final decisions about whether a certain piece of information can be trusted or not may require additional analysis steps. To our knowledge there is no prior work that addresses data trustworthiness in social networks.

The Approach. The approach that we have developed is based on comparing the information in the social network of interest with anonymized data from other social networks (called reference social networks). We start by identifying relevant assumptions. We then introduce our trust score formulation models.

Assumptions. We assume that an organization has created its own social network. Since this network is usually obtained from its own users that willingly share their data with the organization, we call such an organization *data owner*. We use the term *local social network* to refer to the company-owned network. We model this local social network as a graph G = (N, E), where N is the set of nodes and E is the set of edges. Each node represents an individual entity such as a user and each edge represents an existing relation between two nodes. Each node has an associated profile represented by a set of attributes. This set of attribute contains *identifier*, *quasi-identifier*, and occasionally *sensitive* attributes that are supposed to be known by the data owner. We assume that all relationships in this local social network are binary. Moreover, we represent all relationships via unlabeled undirected edges. We use *X* or *Y* to represent individual nodes, and X_i, $i = 1\dots n$, to represent all the nodes in N, where $n = |N|$. We use the notation *X.A* to refer to the attribute *A*'s value for the node *X*. We assume that the owner of the local social network has access to one or more anonymized reference social networks. An anonymized reference social network is provided by an external organization (such as Facebook or LinkedIn) that protects the identity and the sensitive information in the social network data by using an anonymization process. We assume that there are *s* such anonymized reference social networks available. We represent these networks as $AG_j = (AN_j, AE_j)$ ($j = 1, s$). Each such anonymized social network is created by the external organization, owner of the social network, from an original graph. We label the corresponding original graphs as $G_j = (N_j, E_j)$. It is worth noting that these graphs are large compared to the local social network. For simplicity in the presentation, we assume that we have only one target attribute, labeled *B*, which may contain misreported / non-trusted information. When more attributes are non-trusted, we can compute the trust score for one attribute at a time. In order to assess the trustworthiness of values for this attribute *B*, the attribute must exist in each anonymized reference social network (otherwise the anonymized social network is not useful and will not be considered).

Trust Score Formulation. In order to assess the trust score of attribute *B* for node *X* (denoted as *TS(X.B)*) in the local social network, we use all available *s* anonymized reference social networks. To obtain this measure, we use the *intermediary trust score* that we compute for each anonymized reference social network. We use the notation *TSj*(*X.B*) (j = 1 .. *s*) to denote the intermediate trust score obtained by comparing *X.B with* the data in the anonymized reference social network AG*j*. We compute *TSj*(*X.B*), by matching a node *X* from the local social network to nodes from an anonymized reference social network. We consider in this matching, the node attribute's information (that is, the values of attributes *B*, *A*1, .., *Aq*) and the graph structure. The approach used to compute such score is not unique and we plan to investigate three different approaches. The first approach models the trust score as the percentage of nodes from the anonymized reference social network that could potentially be *X*. We refer to this

approach as *absolute trust score* (ATS). The second approach first computes how many nodes from the anonymized network can be *X* when only the trusted attributes *A*1, .., *Aq* and the graph structure are used. We then find the subset of those nodes that match the value of the *B* attribute as well (note that a non-generalized value will match its ancestors on the value generalization hierarchy). The number of those nodes divided by the number of nodes that matches *X* based only on trusted values and graph structure is our second measure of trust. We refer to this measure as *relative trust score* (RTS). Our last approach includes a weight that depends on how the values of attribute *B* are published in the anonymized social network. In most anonymized networks, generalization is used to anonymize the quasi-identifier attributes, and in this case we would like to differentiate between cases when a specific value (such as the exact name of a city) or a generalized value (such as the name of the country) is used. We thus extend the relative trust score computation approach by assigning a higher weight to matches of *X* with anonymized nodes that contain more specific information for *B*. More precisely, the weight associated with a specific value is 1, and the weight decreases when the amount of generalization increases. For example, considering the attribute *city*, the weight associated with a single value like *Chicago* is 1 and the weight associated with a generalized value like *Illinois* is 1/10 assuming that there are 10 cities in Illinois in the value generalization hierarchy used for this attribute. We refer to this approach as *weighted trust score* (WTS).

Once the intermediary trust scores have been obtained for *X.B* with respect to all *s* anonymized reference social networks available, the trust score of *X.B* is defined as the average computed over all these intermediary trust scores. Note that we can use any of the previous approaches for computing the intermediary trust scores and consequently we will obtain three different measures for the resulting trust score.

Preliminary Evaluation. A preliminary experimental evaluation of our trust score model has been carried out. We have used two datasets in the experiments: a synthetic dataset that follows the power-law distribution, generated according to the Albert Barabasi model[1]; and the Enron dataset (available at http://snap.stanford.edu/data). In the experiments, we assumed that reference social networks be anonymized according to the approach by Campan and Truta[2]. Such approach partitions the social network nodes into pairwise disjoint clusters so that any two nodes from any cluster are indistinguishable based on their relationships and quasi-identifier attributes values. The approach uses intra-cluster and inter-cluster edge generalization techniques for generalizing the social network structure. It uses generalization for quasi-identifier attributes values so that each cluster has its profile replaced by the generalization information of that cluster. A key parameter in this approach is represented by the threshold cardinality for each cluster, referred to as parameter *k*.

The experiments compared the recall, that is, the number of detected fake nodes, of the three approaches for computing the intermediary trust score (that is, ATS, RTS, and WTS). In the experiments we varied several parameters, including the value *k* used in the anonymized data sets, the number of fake values, and the magnitude of changes in the fake values compared with the actual data. The results show that all approaches are very effective, especially the WTS approach which has in general a recall greater than 80%. The impact of the magnitude of changes is very interesting; when such magnitude increases the three approaches improve the recall, as the larger the change is in a fake or malicious node, the more easily it is to detect such node. The *k* parameter has also an important impact in that higher values of *k* result in a decrease in recall. The WTS approach is more sensitive to variations in the values of *k*.

[1] A.L. Barabasi, and R. Albert, "Emergence of scaling in social networks", SCIENCE, 286-509, 1999.
[2] A. Campan, and M.T. Truta, "A Clustering Approach for Data and Structural Anonymity in Social Networks", Proc. of PinKDD'08.

XI.B. ASSURING DATA TRUSTWORTHINESS FROM COLLUDING ATTACKS IN SENSOR NETWORKS

New advances in sensor technologies and embedded systems are making possible connecting the physical environment to the Web resulting in the concept of Web-of-things. However for this concept to become a reality, trustworthiness of sensed data must be assured. In this work we have taken a first step towards addressing such requirement by investigating the problem of trustworthiness of data acquired and transmitted by wireless sensor networks (WSNs).

In a WSN, due to a need for robustness of monitoring and low cost of the nodes, there is a certain amount of node redundancy. Data from multiple sensors is aggregated at an aggregator node which then forwards to the base station only the aggregate values. At present, due to limitations of the computing power and energy resource of sensor nodes, data is aggregated by extremely simple algorithms such as averaging. However, such aggregation is known to be very vulnerable to faults, and more importantly, malicious attacks[3]. This cannot be remedied by cryptographic methods, because the attackers generally gain complete access to information stored in the compromised nodes. For that reason data aggregation at the aggregator node has to be accompanied by an assessment of trustworthiness of data from individual sensor nodes. Thus it is critical to be able to estimate the trustworthiness of each sensor node. Thus, more sophisticated algorithms are needed for data aggregation that are able to also assess data trustworthiness and sensor trustworthiness (also referred to as sensor reputation)..

Our approach is based on the use of Iterative Filtering (IF) algorithms[4]. These algorithms are an attractive option for WSNs because they solve both problems- data aggregation and data and sensor trustworthiness assessment- using a single iterative procedure. Such trustworthiness estimate of each sensor is based on the distance of the readings of such a sensor from the estimate of the correct values, obtained in the previous round of iteration by some form of aggregation of the readings of all sensors. Such aggregation is usually a weighted average; sensors whose readings significantly differ from such estimate are assigned less trustworthiness and consequently in the aggregation process in the present round of iteration their readings are given a lower weight. Existing IF algorithms, however, are not robust against sophisticated collusion attacks. In such an attack scenario, colluders attempt to skew the aggregate value by forcing such IF algorithms to converge to skewed values provided by one of the attackers.

Our approach to protect against such attacks is based on providing a robust estimation of errors of individual sensors. When the nature of errors is stochastic, such errors essentially represent an approximation of the error parameters of sensor nodes in WSN such as bias and variance. However, such estimates also prove to be robust in cases when the error is not stochastic but due to coordinated malicious activities. Such initial estimation makes IF algorithms robust against described sophisticated collusion attack, and, we believe, also more robust under significantly more general circumstances; for example, it is also effective in the presence of a complete failure of some of the sensor nodes. Since readings keep streaming into aggregator nodes in WSNs, and since attacks can be very dynamic (such as orchestrated attacks[5]), in order to obtain the trustworthiness scores of sensor nodes as well as to identify compromised nodes we apply our framework on consecutive batches of consecutive readings. Sensors are deemed compromised only relative to a particular batch; this allows our framework to handle on-off type of attacks.

We have carried out extensive experiments, in which we have compared our IF algorithm with other IF algorithms. The results show that whereas the other IF algorithms converge on the skewed values injected

[3] S. Ozdemir and Y. Xiao, "Secure data aggregation in wireless sensor networks: A comprehensive overview," Comput. Netw., vol. 53, no. 12, pp. 2022–2037, Aug. 2009.

[4] C. de Kerchove and P. Van Dooren, "Iterative filtering in reputation systems," SIAM J. Matrix Anal. Appl., vol. 31, no. 4, pp. 1812–1834, Mar. 2010.

[5] K. Hoffman, D. Zage, and C. Nita-Rotaru, "A survey of attack and defense techniques for reputation systems," ACM Comput. Surv., vol. 42, no. 1, pp. 1:1–1:31, Dec. 2009.

by the attackers, our algorithm provides values that are very close approximations of the correct values. The experiments also show that our algorithm converges faster than the other IF algorithms.

XI.C DATA SECURITY IN THE CLOUD

Many benefits, including on-demand provisioning that enables organizations to grow efficiently and in a cost effective manner, have been the driving force of many organizations moving into the cloud. Data as a service (DaaS) is an emerging cloud service where organizations can seamlessly store in the cloud and retrieve based on the access control policies that cover the legal requirements and organizational policies Amazon S3 and Microsoft Azure storage service are two such popular services currently available. An expressive access control model, such as XACML, allows one to specify access control policies (ACPs) on protected objects in terms of the properties of subjects, referred to as identity attributes. The email address, the role a user plays in her organization, the age and the location a user accesses from are a few examples of such identity attributes. The identity attributes that subjects should possess in order to access protected objects are referred to as conditions. Such an attribute-based access control model is very important to provide fine-grained access to data that can easily express policies closer to organizational policies. A crucial issue, often neglected, in this context is that the identity attributes in the conditions often encodes privacy-sensitive information. Many existing cloud data services do provide similar access control models. However, the privacy of the users is not protected in such models. Privacy, both individual as well as organizational, is considered a key requirement in all solutions, including cloud services, for digital identity management. Further, insider threats are considered one of the major source of data theft and privacy breaches. With cloud computing initiatives the scope of insider threats is no longer limited to the organizational perimeter. Therefore, there is a timely need to protect the identity attributes of the users while enforcing attribute-based access control the cloud.

We have developed an approach, called CloudMask[6], based on fine-grained encryption, by which identity attribute access control policies, expressed in a language like XACML, are enforced for data stored by the cloud without the cloud provider learning the values of neither these identity attributes nor any information about the contents of the data.

The approach. The approach has four parties:

- Data Manager (DM): it is an entity at the client organization; this party manages subscriptions and performs policy based encryption of data.
- Cloud Data Service (CDS): is a cloud service hosting the encrypted data.
- Users (Usrs) are the users of the client organization. They register with the DM and retrieve data from the CDS.
- Identity Providers (IdPs) are entities that issue certified identity tokens, i.e., commitments 1 of identity attributes, to Usrs. The IdPs can be part of the client organization or be part of the Proxy (if the client organization trusts the Proxies).

The main idea underlying our approach is based on a fine-grained encryption of the data to be stored in the CDS based on the attribute-based access control (ABAC) policies. By fine-grained encryption we mean that different portions of a data set are encrypted with different keys, depending on the ABAC policies specified for the data; users then receive the keys only for the data they are authorized to access. To address the problem of key distribution, keys are not however directly distributed to the users; rather, each user receives one or more secrets, depending on her identity attributes, by using which she is able to

[6] M.Nabeel, E. Bertino, "CloudMask: Private Access Control in the Cloud" Technical Report, November 2010.

extract the encryption keys. These encryption keys are "hidden'" in a special structure; such a structure is constructed in such a way that any party which does not have the user secret is unable to extract the keys. Therefore the structure does not have to be kept secret and can be stored at the CDS or at a web site or broadcasted on an unsecure channel. The approach is based on two building block:

- Oblivious Commitment Based Envelope (OCBE) protocols. These protocols provide a way to obliviously deliver a message to the Usrs who satisfy certain conditions. The DM and Usrs engage in OCBE protocols for Usrs to obtain secrets for their identity tokens, expressed as commitments. For a given condition c, a Usr sends her identity token, obtained from an IdP, to the DM. The DM, in turn, sends the Usr an envelope, that is, an encrypted message, containing a secret. The Usr can open (i.e., decrypt) the envelope only if she knows the committed value in her identity token. In other words, the Usr can derive the symmetric key only if her identity token verifies the condition c.
- Group Key Management (BGKM) schemes. A novel BGKM scheme based on the really simple idea of matrix null spaces is used[7]. Such an approach ensures forward and backward secrecy and only requires to store together with the encrypted data an access control vector. This vector does not need to be encrypted, as only the users who have certain secrets can use the vector to extract the encryption key. An important advantage of this approach is that if new keys have to issued, the client organization just needs to send to the cloud the newly encrypted data and the new access control vector; however it does not need to send any information to the users, as the users can simply extract the new key from the new access control vector.

XII. IDENTIFICATION OF RELATED INFORMATION OF INTEREST ACROSS FREE TEXT LAW ENFORCEMENT DOCUMENTS

James Johnson and Anita Miller, ADB Consulting
Latifur Khan, Collaborator, The University of Texas at Dallas

The goal of the project was to explore semantic processing approaches for determining related information of interest between documents on a sub-sentence level. As the research progressed it was discovered that semantic graph matching methods yield both related information of interest as well as new augmented information linked to the graph being matched to a reference graph. It was also found that relatedness and augmented information measures could be used to sort relevant information. It became clear that the developed approach is applicable to other domains such as intelligence, monitoring of news reports, identifying cyber threats and attacks, as well as information of interest-driven internet searches. The approach was tested with encouraging results against cyber threat messages exchanged by the Anonymous hactivists and against published FBI reports.

This research significantly expanded semantic analysis of free text by 1) quantifying semantic content and semantic context, 2) incorporating DLSafe Rules and abductive hypotheses that model processes and generate inferences for increased likelihood of matching related content and discovering new knowledge, and 3) creating a rigorous definition of a new expanded semantic graph structure with semantic relatedness measures for quantifying identified information on a level not previously achieved. These new techniques lay the foundation for cross-domain applications including the support of national intelligence analysts who need to identify focused information from large volumes of free text.

[7] N. Shang, M. Nabeel, F. Paci, and E. Bertino, "A privacy-preserving approach to policy-based content dissemination," in ICDE '10: Proceedings of the 2010 IEEE 26th International Conference on Data Engineering, 2010.

Relatedness measures were developed for semantic content (entity phrases plus associated attributes), semantic context (links between semantic content as well as inferred information). These measures identified degrees of related information between multiple free-text documents. Attributes were assigned importance values based on statistical feature distributions (results verified with participating law enforcement investigators). In addition, a semantic augmentation measure, defined on an expanded semantic graph, measured the relevance of information connected to the compared graph using the significance of the information associated with each additional connected edge. This augmented information revealed new leads for the investigators. The measures enabled sorting of the results so that the most relevant leads could be presented first.

The evaluation of the developed research was conducted on over 310,000 comparisons of emails exchanged between law enforcement investigators across a large geographical area and collected over a four year period. The data characteristics included a general lack of punctuation, liberal use of capitalization, domain-specific terminology, acronyms, abbreviations and slang. They also contained cut–and-paste insertions and attachments. The evaluation showed that the new relatedness measures were effective and valuable. Of particularly notable interest were the large number of times that related information of interest was successfully found, how well unimportant related information was eliminated, and how frequently augmented knowledge was additionally identified.

Additional related research topics have been identified as a result of this research that could further the understanding of free text semantic processing techniques, provide a basis for information theory applications, and enable the extraction of event threads from expanded semantic graphs.

Three papers were published and presented during 2012. In addition a keynote address was presented on “Detecting Emergent Terrorist Events: Finding Needles in Haystacks” at the 2012 European International Security and Informatics Conference, and a presentation was made on “Identification of Cyber Threats” at the Texas Security Week held at The University of Texas at Dallas.

PART B: SAMPLE PUBLICATIONS AND PRESENTATIONS

Books:

Data Mining Tools for Malware Detection (partially supported by the project)
Mehedy Masud, Latifur Khan and Bhavani Thuraisingam
Taylor and Francis, December 2011

Developing and Securing the Cloud (partially supported by the project)
Bhavani Thuraisingham, CRC Press, November 2013

Secure Data Provenance and Inference Control (partially supported by the project)
Tyrone Cadenhead, Murat Kantarcioglu, Bhavani Thuraisingham, Vaibhav Khadilkar
Book completed, gone through editorial and is now with the printer (expected publication Sept 2014)

Analyzing and Securing Social Media
Satyen Abrol, Raymond Heatherly, Latifur Khan, Murat Kantarcioglu, Vaibhav Khadilkar, Bhavani Thuraisingham, Contract Signed with Taylor and Francis), Several Chapters Completed, Publication 2015.

Digital National Security and Cyber Operations
Jan Kallberg and Bhavani Thuraisingham
Contract Signed with Taylor and Francis

Journal Publications

Masud, M. M., Al-Khateeb, T. M., Hamlen, K., Gao, J., Khan, L., Han, J. and Thuraisingham, B. Cloud-based Malware Detection for Evolving Data Streams. ACM *Transactions on Management Information Systems*, 2011.

Mohammad M. Masud, Clay Woolam, Jing Gao, Latifur Khan, Jiawei Han, Kevin Hamlen, and Nikunj C. Oza. Facing the reality of data stream classification: Coping with scarcity of labeled data. *Journal of Knowledge and Information Systems (KAIS)*, 2011.

Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani M. Thuraisingham: Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints. *IEEE Trans. Knowl. Data Eng.* 23(6): 859-874 (2011)

Mohsen Rezvani, Aleksander Ignjatovic, Elisa Bertino, and Somesh Jha: Secure data aggregation technique for wireless sensor networks in the presence of collusion attacks. To appear in IEEE Transactions on Dependable and Secure Computing.

Mohamed Nabeel, Ning Shang, Elisa Bertino: Privacy Preserving Policy-Based Content Sharing in Public Clouds. IEEE Trans. Knowl. Data Eng. 25(11): 2602-2614 (2013)

Pallabi Parveen, Nathan McDaniel, Zackary R. Weger, Jonathan Evans, Bhavani M. Thuraisingham, Kevin W. Hamlen, Latifur Khan: Evolving Insider Threat Detection Stream Mining Perspective. International Journal on Artificial Intelligence Tools 22(5) (2013)

Satyen Abrol, Latifur Khan, Fahad Bin Muhaya: MapIt: a case study for location driven knowledge discovery and mining. IJDMMM 5(1): 57-75 (2013)

Conference Publications

Abhijith Shastry, Murat Kantarcioglu, Yan Zhou, and Bhavani Thuraisingham, "Randomizing smartphone malware profiles against statistical mining techniques" in *Proceedings of the 26th Annual IFIP WG 11.3 conference on Data and Applications Security and Privacy* (DBSec'12), Nora Cuppens-Boulahia, Frédéric Cuppens, and Joaquin Garcia-Alfaro (Eds.). Springer-Verlag, Berlin, Heidelberg, 239-254.

Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Bowei Xi, "Adversarial support vector machine learning" in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '12). ACM, New York, NY, USA, 1059-1067.

Yan Zhou, Murat Kantarcioglu, and Bhavani Thuraisingham, "Sparse Bayesian Adversarial Learning Using Relevance Vector Machine Ensembles" to appear in *Proceedings of the 12th IEEE International Conference on Data Mining, ICDM 2012*

Satyen Abrol, Latifur Khan and Bhavani Thuraisingham,"*Tweeque: Spatio-Temporal Analysis of Social Networks for Location Mining Using Graph Partitioning*," The First ASE/IEEE International Conference on Social Informatics, December 14-16, 2012, Washington D.C., USA.

Satyen Abrol, Latifur Khan, Vaibhav Khadilkar, Bhavani M. Thuraisingham, Tyrone Cadenhead: Design and implementation of SNODSOC: Novel class detection for social network analysis. ISI 2012: 215-220

Satyen Abrol, Latifur Khan and Bhavani Thuraisingham, "*Tweecalization: Efficient and Intelligent location mining in Twitter using semi- supervised learning*," 8th IEEE International Conference on Collaborative Computing, October 14–17, 2012, Pittsburgh, Pennsylvania

Satyen Abrol, Latifur Khan, Bhavani M. Thuraisingham: Tweeque: Spatio-Temporal Analysis of Social Networks for Location Mining Using Graph Partitioning. Social Informatics 2012: 145-148

Hyo-Sang Lim, Gabriel Ghinita, Elisa Bertino, Murat Kantarcioglu: A Game-Theoretic Approach for High-Assurance of Data Trustworthiness in Sensor Networks. ICDE 2012: 1192-1203

J. Johnson, A. Miller, L. Khan, and B. Thuraisingham, "Extracting semantic information structures from free text law enforcement data," *IEEE Intelligence and Security Informatics (ISI),* Washington, D.C., July 11-14, 2012.

J. Johnson, A. Miller, L. Khan, and B. Thuraisingham, "Measuring Relatedness and Augmentation of Information of Interest within Free Text Law Enforcement Documents," *2012 IEEE European International Security and Informatics, (EISI),* Odense, Denmark, Aug 22-24, 2012.

J. Johnson, A. Miller, L. Khan, B. Thuraisingham, "Graphical Representation of Semantic Information," *International Conference on Semantic Computing (ICSC),* Palermo, September 19-21, 2012.

Jan Kallberg: The Common Criteria Meets Realpolitik: Trust, Alliances, and Potential Betrayal. IEEE Security & Privacy 10(4): 50-53 (2012)

Jan Kallberg, Bhavani M. Thuraisingham: Towards cyber operations - The new role of academic cyber security research and education. ISI 2012: 132-134

Pallabi Parveen, Bhavani M. Thuraisingham: Unsupervised incremental sequence learning for insider threat detection. ISI 2012: 141-143

Mohsen Rezvani, Aleksandar Ignjatovic, Elisa Bertino, Sanjay Jha: A robust iterative filtering technique for wireless sensor networks in the presence of malicious attacks. Poster Paper, ACM 2013 SenSys 2013: 30.

Chenyun Dai, Fang-Yu Rao, Traian Marius Truta, Elisa Bertino: Privacy-preserving assessment of social network data trustworthiness. CollaborateCom 2012: 97-106 (extended version invited for publication in International Journal of Cooperative Information Systems).

Hyo-Sang Lim, Gabriel Ghinita, Elisa Bertino, Murat Kantarcioglu: A Game-Theoretic Approach for High-Assurance of Data Trustworthiness in Sensor Networks. ICDE 2012: 1192-1203.

Tyrone Cadenhead, Murat Kantarcioglu, Vaibhav Khadilkar, Bhavani M. Thuraisingham: Design and Implementation of a Cloud-Based Assured Information Sharing System. MMM-ACNS 2012: 36-50

Bhavani M. Thuraisingham, Vaibhav Khadilkar, Jyothsna Rachapalli, Tyrone Cadenhead, Murat Kantarcioglu, Kevin W. Hamlen, Latifur Khan, Mohammad Farhan Husain: Cloud-Centric Assured Information Sharing. PAISI 2012: 1-26

Tyrone Cadenhead, Vaibhav Khadilkar, Murat Kantarcioglu, Bhavani M. Thuraisingham: A cloud-based RDF policy engine for assured information sharing. SACMAT 2012: 113-116

Tyrone Cadenhead, Vaibhav Khadilkar, Murat Kantarcioglu, Bhavani M. Thuraisingham: A language for provenance access control. CODASPY 2011: 133-144

Tyrone Cadenhead, Vaibhav Khadilkar, Murat Kantarcioglu, Bhavani M. Thuraisingham: Transforming provenance using redaction. SACMAT 2011: 93-102

Jyothsna Rachapalli, Vaibhav Khadilkar, Murat Kantarcioglu, Bhavani M. Thuraisingham: REDACT: a framework for sanitizing RDF data. WWW (Companion Volume) 2013: 157-158

Daniel C. Krawczyk, James Bartlett, Murat Kantarcioglu, Kevin W. Hamlen, Bhavani M. Thuraisingham: Measuring expertise and bias in cyber security using cognitive and neuroscience approaches. ISI 2013: 364-367

Jan Kallberg, Bhavani M. Thuraisingham: Towards cyber operations - The new role of academic cyber security research and education. ISI 2012: 132-134

Jan Kallberg, Bhavani M. Thuraisingham, Erik Lakomaa: Societal Cyberwar Theory Applied: The Disruptive Power of State Actor Aggression for Public Sector Information Security. EISIC 2013: 212-215

Pallabi Parveen, Pratik Desai, Bhavani M. Thuraisingham, Latifur Khan: MapReduce-guided scalable compressed dictionary construction for evolving repetitive sequence streams. CollaborateCom 2013: 345-352

Pallabi Parveen, Nathan McDaniel, Zackary R. Weger, Jonathan Evans, Bhavani M. Thuraisingham, Kevin W. Hamlen, Latifur Khan: Evolving Insider Threat Detection Stream Mining Perspective. International Journal on Artificial Intelligence Tools 22(5) (2013)

Pallabi Parveen, Nate McDaniel, Varun S. Hariharan, Bhavani M. Thuraisingham, Latifur Khan: Unsupervised Ensemble Based Learning for Insider Threat Detection. SocialCom/PASSAT 2012: 718-727

Pallabi Parveen, Bhavani M. Thuraisingham: Unsupervised incremental sequence learning for insider threat detection. ISI 2012: 141-143

Keynote Presentations:

Several keynote addresses have been given on our research on cloud-centric assured information sharing including the following:

1. *Assured Cloud-based Information Sharing*, IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC), December 2011, Sydney, Australia.
2. *Assured Cloud Computing*, Cyber Security Conference, Cyber security conference at Arizona State University, April 2012
3. *Assured Cloud-based Information Sharing*, PAISI (Pacific Asia Intelligence and Security Informatics) May 2012, Kuala Lumpur, Malaysia. (presented by Prof. Latifur Khan)
4. *Assured Cloud-based Information Sharing*, International Symposium on Foundation of Open Source Intelligence and Security Informatics (FOSINT 2012), August 2012, Istanbul Turkey.
5. *Assured Cloud-based Information Sharing*, AFOSR-EOARD Conference (Intl. Conference on Mathematical Models, Models and Architectures for Computer Network Security), October 2012, St. Petersburg, Russia (presented by Latifur Khan).
6. *Assured Cloud-based Social Networking*, Chinese Academy of Sciences Conference on Social Computing, Beijing, China, November 2012.
7. *Secure Cloud Computing*, University of North Texas/Collin College SoMiC Workshop on Cyber Security, Denton, TX, April 2013.
8. Analyzing and Securing Social Networks, WWW Workshop on Social Network Security and Privacy, Rio De Janeiro, Brazil, May 2013.
9. *Cloud-based Assured Information Sharing*, Conference on Security, Privacy and Trust, Melbourne Australia, July 2013.
10. *Analyzing and Securing Social Networks*, Society for Design and Process Science World Conference, Campenas, Brazil, October 2013.
11. *Cloud-based Assured Information Sharing*, IEEE Cloudcom, Bristol UK, December 2013.
12. Cloud-Centric Assured Information Sharing, IEEE Workshop on Big Data Security and Privacy, Anchorage Alaska, June 2014.

PART C: TOOL REPOSITORIES

UTDallas Data Mining Tool Repository (Partially supported by the project) Latifur Khan

We currently maintain a number of datasets along with the tools that implement our algorithms as a service to the data mining and security communities. This web page is intended to facilitate researchers' and developers' access to the source code of our algorithms and datasets rather than developing them from scratch. Users can download the source code and run it for our datasets and/or for their own datasets. In addition, they can extend their research by utilizing our software. As of today, this web page describes a number of tools. The first tool implements malware code detection; the second tool describes novel class detection for stream data; and the third describes stream data classification with limited labeled data. We will be adding more tools (e.g., Code Blocker, Data Mining for Ontology Alignment) as they become available. Please see the following link for more details: http://dml.utdallas.edu/Mehedy/index.html

UTDallas Secure Cloud Tool Repository (Partially supported by the project) Bhavani Thuraisingham

We have set up a secure cloud repository and will continue to enhance this repository. Details can be found at http://cs.utdallas.edu/secure-cloud-repository/

PART D: STUDENTAND POSTDOCS SUPPORTED

Ph.D. Students

Wei-She (partial support, graduated December 2011 – Now at Intel Corporation)

Parveen Pallabi – graduated December 2013 – Now at VCE Corporation

Vaibhav Khadilkar (partial support, graduated December 2013, now at NutraSpace)

Jyothsna Rachpalli (partial support))

Satyen Abrol (Graduated May 2013, Now at VMWare)

Lidan Fan (Graduated May 2014, Starting faculty position at U of TX at Tyler)

Mohamed Nabeel (Purdue, Graduated 2013, now at Oracle Corp).

Postdocs/Professional Staff/Masters Students

Nathan McDaniel (Professional Software Developer, October 2012 – May 2014)

Yan Zhou (Partial support – Postdoc September 2010 - August 2012)

Jan Kallberg (Postdoc – July 2012 – December 2013)

Mehedy Masud (Postdoc – March 2010 - December 2011)

Tyrone Cadenhead (Postdoc – September 2011 – June 2013)

Rhonda Walls, Project Coordinator (Partial Support, June 2012 - May2014)

Guna Rajasekar (Masters Student, January 2014 – May 2014)

Gautum Ganesh (Masters Student, January 2014 – May 2014)

PART E: PATENTS FILED AND TEHNOLOGY TRANSFER

1. US Patent Application No: 2012/0054,184 Systems and Methods for Detecting a Novel Data Class, Mehedy Masud, Latifur Khan, Bhavani Thuraisingham, Jiawei Han et al. (Partially supported by the project)

2. US Application No. 13/588,977 Systems And Methods For Determining User Attribute Values By Mining User Network Data And Information" Satyen Abrol, Latifur Khan, Bhavani Thuraisingham, Vaibhav Khadilkar

3. US Serial Number 62/015678 SYSTEM AND METHOD FOR THREAT DETECTION AND PREDICTION (Marketing)**,** Satyen Abrol, Latifyr Khan, Vaibhav Khadilkar, Bhavani Thuraisingham, Nathan McDaniel

4. **US Serial No.: 62/015,697**SYSTEM AND METHOD FOR THREAT DETECTION AND PREDICTION (Security),
G. Rajaseker, G. Ganesh, L. Kjan, B. Thuraisingham, N. McDaniel, S. Abrol, V. Khadilkar

Patent applications 3 and 4 are on systems developed based on patents 1 and 2. We are exploring ways to commercialize the systems through UTD's office of technology commercialization. Descriptions of these systems are as follows:

Threat Detection and Prediction

Like a blunt instrument which destroys more than is intended, NSA's (National Security Agency) PRISM program dredges the communications landscape and gathers more information than should be necessary to ferret out terrorists and terrorist cells communicating inside the US (United States) and around the world. The NSA PRISM program is deemed necessary in order to prevent future terrorist acts against the US. This top-down approach not only breaches the privacy of US citizens and upset and angered them but it has also drawn the ire of foreign governments who have been spied upon.

By contrast, InXite uses a bottom-up approach that uses specific keywords designed to reveal people around the world tweeting about a topic of particular interest. For instance the keyword pair "Egypt" and "Muslim-brotherhood" would display a list of people in Egypt tweeting to others around the world using the keyword "Muslim-brotherhood". In other words InXite uses a targeted approach without needing to gather massive amounts of data.

In addition InXite integrates information from a variety of online social media sites such as Foursquare, Google+ and LinkedIn, builds people profiles through correlation, aggregation and analyses in order to identify persons of interest who pose a threat. Other applications include garnering user feedback on a company's products, providing inexpensive targeted advertising, and monitoring the spread of an epidemic, among others.

This invention describes our cloud-based system InXite, also called InXite-Security (Stream-based Data Analytics for Threat Detection and Prediction), that is designed to detect evolving patterns and trends in streaming data including emails, blogs, sensor data and social media data. InXite is designed on top of two powerful data mining systems, namely Tweethood (location extraction for Tweets), with the explicit aim of detecting and predicting suspicious events and people and SNOD (Stream-based Novel Class Detection). We also designed a separate system, SNOD++, an extension of SNOD, for detecting multiple novel classes of threats for InXite. Our goal is to decipher and monitor topics in data streams as well as to detect when trends emerge. This includes general changes in topics such as sports or politics and also includes new, quickly-emerging trends such as hurricanes and bombings. The problem of correctly associating data streams (*e.g.,* Tweet messages) with trends and topics is a challenging one. The

challenge is best addressed with a streaming model due to the continuous and large volume of incoming messages.

Data streams are emanating from numerous data sources including blogs and social media data. Such data could be structured, unstructured, semi-structured, and real-time/non real-time, static or dynamic data. It also includes relational and object data as well as semantic web data such as Resource Description Framework (RDF) graphs and multimedia data such as video, audio, and images. With modern technology, it is possible to exchange numerous messages in a very short space of time. Furthermore, communication messages (*e.g.*, blogs and tweets) are often abbreviated and difficult to follow. To understand the motives, sentiments and behavior of individuals and groups, where some of them could be malicious, tools are needed to make sense out of the massive amounts of streaming data often represented as graphs. To address this need, we have designed a framework called InXite for analyzing stream-based data.

We have utilized Tweethood and SNOD to develop a sophisticated data analytics system called InXite. InXite is a multi-purpose system that can be applied to security marketing, healthcare and financial applications among others. We have designed and developed two InXite applications. One is InXite-Security and the other is InXite-Marketing. InXite-Security (which we will refer to as InXite for convenience since much of the InXite system initially focused on security applications) will detect and predict threats including potential terrorists, harmful events and the time and place of such events. The data sources for InXite include blogs, sensor and social media data amongst others. InXite is a cloud-based application due to the numerous advantages of clouds such as on-demand scalability, reliability and performance improvements. InXite-Marketing utilizes the various modules of InXite and gives recommendations to businesses for selling products. The design of InXite uses Tweethood to obtain demographics information about individuals and SNOD and SNOD++ for detecting novel classes of threats and sentiments.

InXite-Marketing: Media Data Analytics

This invention describes a real-time, cloud-based social media analytics system called InXite-Marketing that is designed to identify potential customers to target based on the micro-level location of the users and their behavior observed from their Twitter activity. It builds on our previous invention- “InXite-Security” which deals with threat detection and prediction for security applications.

InXite-Security utilizes two patent pending technologies called Tweethood and SNOD to develop a sophisticated data analytics system. The backbone of InXite-Security, which is called InXite, is a multi-purpose system that can be applied to security, marketing, healthcare and financial applications among others. InXite integrates data from multiple social networks as well as databases and carries out analytics. InXite-Security is built on InXite and detects and predicts threats including potential terrorists, harmful events and the time and place of such events. The data sources for InXite include blogs, sensor and social media data amongst others. Our invention disclosure for InXite-Security includes a discussion of the InXite framework and described the modules we have developed for threat detection and prediction. InXite-Marketing utilizes the foundational technologies utilized by InXite-Security and in addition has several innovative modules for social media based marketing. It is a real-time social data analytics system, which gives instant insights to businesses for selling their products and services.

Two of the major modules of InXite-Marketing are Sentiment Analysis and Recommender Systems. The sentiment analysis module predicts the sentiment for a person on a subject based on his/her tweets. For example, if we want to determine John Smith's sentiment about the iPhone-5 - we can go through all his tweets about the iPhone-5 and based on the sentiment of each of the tweets, we can determine the overall sentiment of John Smith about iPhone-5. Once we know the sentiment of John Smith about iPhone-5, say it is positive, we can recommend to him some more i-products or products related to the iPhone-5 such as a headphone or a charger.

F. EDUCATION

We have developed new courses as well as enhanced several of our existing courses based on the research carried out for this project.

New Courses

1. Developing and Securing the Cloud (Spring 2012, Spring 2014)
2. Analyzing and Securing Social Media (Spring 2013)
3. Big Data Analytics (Fall 2012, Spring 2013, Fall 2013, Spring 2014)

Enhancements to Existing Courses

Data and Applicators Security (Fall 2010, Fall 2011, Fall 2012, Fall 2013)

Cyber Security Essentials (Summer 2011, Summer 2012, Summer 2013)

Partly due to the success of the AFOSR project:

We are preparing courses for IBM under an IBM Faculty Award

Received an NSF Capacity Building Grant for Assured Cloud Computing to develop courses in assured cloud computing.

3 Directions

We have made substantial progress in research, education and technology transfer. Future directions will include the following.

Research: We will continue to write research proposals to NSF as well as to the various agencies to conduct fundamental research on integrating secure social media, mobile computing and cloud computing.

Prepare white papers for 6-2 and 6-3 funding.

Education: Continue with enhancing courses and adding new courses.

Technology Transfer: Explore opportunities to commercialize InXite.